Pergamon

Computers Math. Applic. Vol. 32, No. 5, pp. 99–115, 1996
Copyright©1996 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0898-1221/96 $15.00 + 0.00

S0898-1221(96)00138-1

Polynomial Approximation of Images

I. SADEH*
Department of Communications Engineering
Holon Institute of Technology, 52 Golomb Street, Holon, Israel
sade@math.tau.ac.il

(Received December 1995; accepted February 1996)

Abstract—A transform method for compressing digital data on a two-dimensional lattice, based on polynomial approximation is studied. The least square polynomial approximation in the sense that the "distance" between the original and the reconstructed image is used. The mathematical results establish a basis for a useful implementation for image compression, and transmission having a specific tolerable distortion. Properties of the set of matrix-compressors which minimizes the distance, the matrix-transform, and the minimal error are explicitly found. A necessary and sufficient condition for "distortion-free compression" (i.e., lossless-compression) is found. The tradeoff between the image distortion and the transmitted information is discussed. Experimental results address practical issues of quantization, bit allocation, pyramidal decomposition, and error distribution.

Keywords—Image compression, Polynomial approximation, Transform method, Pyramidal structure.

1. INTRODUCTION

Digital representation of images requires a very large number of bits. It is highly important in almost all practical applications to represent the image, or the information contained in the image, with fewer bits. Recent advances in technology have made it practical to store and communicate high bandwidth analog data, such as images and video in a digital form. Digital communication provides flexibility, reliability, and cost effectiveness, with the added advantage of communication privacy and security through encryption. Digital representation of images allows us to process the data more efficiently and effectively.

The key obstacle for many applications is the vast amount of data required to represent a digital image directly. A digitized version of a single, color picture at TV resolution contains of the order of one million bytes; 35 mm film resolution requires ten times that amount. The use of digital images often is not viable due to high storage or transmission costs, even when image capture and display devices are quite affordable.

Lossless data compression works well for textual data, but its performance in case of digitized data is rather limited. Lossy compression, on the other hand, can be designed to offer a wide range of compression rates at the expense of the quality of the reconstructed image. State-of-the-art techniques can compress typical images from 1/10 to 1/50 their uncompressed size, without visibly affecting image quality. The issue now becomes, how to achieve a certain better

*This research is supported by the Ministry of Science of Israel by Eshkol Fellowship No. 0375. The author is grateful to I. Dinstein for his useful advice. Thanks to M. Shkliarman and V. Avrin for providing the figures in Section 5.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

compression ratio while optimizing the signal quality, or vice versa, how to maintain a desired quality while minimizing the data rate. However, compression is always achieved at the cost of computational resources. Source coding with a fidelity criterion (also called rate distortion theory) is a field of information theory originated by Shannon about 35 years ago [1]. The reader interested in source coding theory can consult one of the texts [2–4].

In this paper, we deal with a theory that emerges from approximation theory rather than information theory. We shall focus upon theory, rather than practice. The reader interested in applications can profitably consult the references [5–9] for the treatment of practical implementation issues not discussed in this paper.

We study algorithms based on classical approximation and interpolation methods, some of them were experimentally studied in [7]. Our contribution lies on new theoretical results related to issues raised in [7], and some other experimental complimentary aspects. We assume that the image has already been segmented into a set of regions, such that region boundaries fit contours. However, in practice the region boundaries should be modified recursively as explained in the description of the adaptive split-and-merge coding method [7]. We shall not discuss such adaptive methods in this paper. Assuming that a segmentation fits region borders to the contours in the scene, it is most common to observe that the image data is a slowly varying luminance function over any region. Such variations are very well represented by 2-D polynomials. Moreover, the oscillatory behavior of commonly used orthogonal functions, difficult to define over nonrectangular regions, do not exist for polynomials. The discussion outlined in [7] is presented here to give a clear motivation for the polynomial approximation, cost function, and optimal compression as used in this paper. We shall prove properties of the proposed transform and that the matrix-transform is not unique. It is shown, that there is a convex set of such matrices, and we present a method for generating all the members of the set. Thus, it is useful to search in the convex set of all "compressors" for the one which is best for encoding. It is not clear that the minimal absolute value element in the set is also the best for image representation, as is done in [7,10]. However, we used the minimum absolute value matrix in our experiments.

Moreover, we find a necessary and sufficient condition of lossless compression, and in general for a fixed LSE distortion. Later, we define a new cost function which is a tradeoff of the error and the amount of transmitted information. Such a combination is useful for adaptive systems where cost and performance parameters should be controlled in almost real time. We discuss briefly such methods.

As explained in Burt and Adelson [11] and Mallat [12], the "secret" of a good compression technique is to transmit a low-pass filtered image at a low rate, and the "error" which is decorrelated needs many fewer bits. But, we believe that the low-pass filtered image should be calculated such that under the constraint of the required rate, it will be optimal in the "best-approximation" sense as it is known in the literature (cf., Davis [13]). Since neighboring pixels are highly correlated, then there should be best approximation in L^2 or in other objective fidelity criteria such as "peak value of error."

We represent the image $F(i,j)$ as a sum of the best approximated image $\hat{F}(i,j)$, and the error $E(i,j)$. That is, $F(i,j) = \hat{F}(i,j) + E(i,j)$. The same argument repeats: $\hat{F}(i,j)$ is a low-pass filtered image and may be encoded at a reduced sample rate, and $E(i,j)$ may be represented with fewer bits per pixel than $F(i,j)$. Further data compression is achieved by iterating this process. Using Burt's terminology: $g_0(i,j)$ is the original image, $g_1(i,j)$ is the approximated image, and $L_0(i,j)$ is the error. Then

$$
\begin{aligned}
L_0(i,j) &= g_0(i,j) - g_1(i,j), \\
L_1(i,j) &= g_1(i,j) - g_2(i,j), \\
&\vdots \\
L_n(i,j) &= g_n(i,j) - g_{n+1}(i,j).
\end{aligned}
$$

By repeating these steps several times we obtain a sequence of two-dimensional arrays $L_0, L_1, \ldots, L_n$. In order to simplify the computation, we choose, like Burt, a resolution step equal to 2. The details at each resolution 2^j are calculated by best approximation of the difference of previous "filtered" images, and by subsampling the resulting image by a factor 2^j. So $g_0(i,j)$ has $M \times N$ pixels, and $g_1(i,j)$ is approximated to $M/2$, $N/2$ degrees of polynomials.

Next, we address practical issues encountered in lossy compression of images by polynomial approximation, such as: quantization, bit-rate versus error, segmentation, variance of the matrix coefficients, pyramidal structure of images, and histograms of the error. Experimental results of image compression using those schemes are presented and discussed.

2. DIGITAL IMAGE COMPRESSION

We treat the 2-dimensional case. The n-dimensional case can be treated by the same tools.

The data structure of the problem is a 2-D lattice with K pixels, where the set of the known pixels is composed as a Cartesian product $\mathcal{K} = \underline{k} \times \underline{l}$, where $\dim \underline{k} = M$ and $\dim \underline{l} = N$ (see also [13]). Denote: $\underline{k} \triangleq \{k_1, \ldots, k_M\}$, and $\underline{l} \triangleq \{l_1, \ldots, l_N\}$.

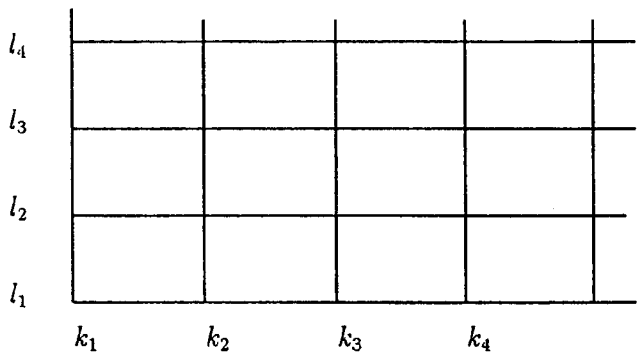

Figure 1. The 2-D lattice $\mathcal{K} = \underline{k} \times \underline{l}$.

For $0 \leq i \leq M$, $0 \leq j \leq N$ let $F(i,j)$ be the sampled digital image on $\mathcal{K}$. The reconstructed image $\hat{F}(i,j)$ is also defined on $\mathcal{K}$. For a variety of considerations outlined in [7], the approximation criterion we have used is the least square error between the original and approximated data. We define the following notations.

(1) M', N': The dimension of the reduced data $M' \leq M$, $N' \leq N$.
(2) A: a matrix of dimension $M' \times N'$.
(3) $\tilde{A}_{M',N'}$: The set of all "compressors" A, that achieve minimal distortion for M', N'.
(4) $I_0 = M' \times N'$: The "amount" of transmitted information.
(5) $d(A) = \|F - \hat{F}\|$: is the Euclidian norm.
(6) The compressor operator is C: $F(i,j)_{M \times N} \rightarrow A(\tilde{i},\tilde{j})_{M' \times N'}$, where $M' \leq M$, and $N' \leq N$.
(7) The decompressor operator is D: $A(\tilde{i},\tilde{j})_{M' \times N'} \rightarrow \hat{F}(i,j)_{M \times N}$, where $\hat{F}(x,y) = \sum_{i=1}^{M'} \sum_{j=1}^{N'} a_{i,j} x^{i-1} y^{j-1}$.

The general scheme is described in Figure 2.

(8) The "reduced Vandermonde" matrices in both dimensions are:

$$V'_{(M')}(\underline{k}) \triangleq \begin{bmatrix} 1 & k_1 & k_1^2 & \cdots & k_1^{M'-1} \\ 1 & k_2 & k_2^2 & \cdots & k_2^{M'-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & k_M & k_M^2 & \cdots & k_M^{M'-1} \end{bmatrix}, \quad k_1 < k_2 < \cdots < k_M, \quad \dim V'_{(M')}(\underline{k}) = M \times M',$$

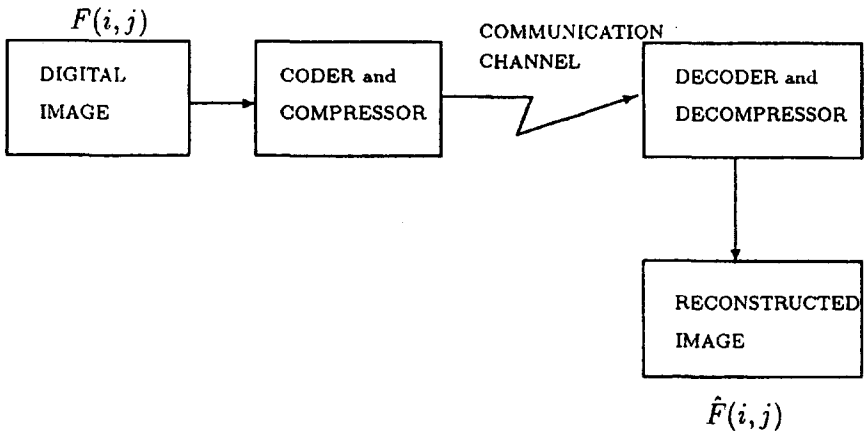

Figure 2. General compression and decompression scheme.

$$V'_{(N')}(\underline{l}) \triangleq \begin{bmatrix} 1 & l_1 & l_1^2 & \cdots & l_1^{N'-1} \\ 1 & l_2 & l_2^2 & \cdots & l_2^{N'-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & l_N & l_N^2 & \cdots & l_N^{N'-1} \end{bmatrix}, \quad l_1 < l_2 < \cdots < l_N, \quad \dim V'_{(N')}(\underline{l}) = N \times N'.$$

Obviously, each pixel at the reconstructed image $\hat{F}$ is defined as:

$$\hat{F}(k,l) = \sum_{i=1}^{M'} \sum_{j=1}^{N'} a_{i,j} k^{i-1} l^{j-1}, \qquad (k,l) \in \mathcal{K}, \quad K = |M \times N| = |\mathcal{K}|.$$

We describe these MN equations in a matrix form of dimension $M \times N$

$$\hat{F}(i,j) = \begin{bmatrix} 1 & k_1 & k_1^2 & \cdots & k_1^{M'-1} \\ 1 & k_2 & k_2^2 & \cdots & k_2^{M'-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & k_M & k_M^2 & \cdots & k_M^{M'-1} \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1,N'} \\ a_{12} & \cdots & a_{2,N'} \\ \vdots & \cdots & \vdots \\ a_{M',1} & \cdots & a_{M',N'} \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \\ l_1 & \cdots & l_N \\ l_1^2 & \cdots & l_N^2 \\ \vdots & \cdots & \vdots \\ l_1^{N'-1} & \cdots & l_N^{N'-1} \end{bmatrix}.$$

Hence,

$$\hat{F}(\underline{k},\underline{l}) = V'_{(M')}(\underline{k})\, A V'_{(N')}(\underline{l})^{\top},$$

where A is the optimal compressor $A \in \tilde{A}(M'N')$.

During the compression process (the approximation method), we use the basic characteristic of images that neighboring pixels are highly correlated. By preserving this structure we enable 2-D approximation. For convenience, we will assume that the rectangular region defined by the segmentation fits to the contours of objects in the scene. It is possible to compress the image in either the vertical or the horizontal direction by reducing the polynomial order. The preferred direction should be chosen according to the characteristics of the image.

First we study the properties of the "reduced Vandermonde matrices."

LEMMA 1. *$V'_{(M')}(\underline{k})$ and $V'_{(N')}(\underline{l})$ have the following properties:*

(1) *$|V'_{(M')}(\underline{k})| > 0$ and $|V'_{(N')}(\underline{l})| > 0$, for all $\underline{k}$, $\underline{l}$, $M' \leq N$, $N' \leq N$, and*
(2) *rank $V'_{(M')}(\underline{k}) = M'$ and rank $V'_{(N')}(\underline{l}) = N'$, for all $\underline{k}$, $\underline{l}$, $M' \leq N$, $N' \leq N$.*

PROOF. The proof for $V'_{(M')}(\underline{k})$ is identical to $V'_{(N')}(\underline{l})$, and therefore, will be omitted.

(1) From the definition of $V'_{(M')}(\underline{k})$, where all the k_i are distinct, therefore, $|V'_{(M')}(\underline{k})| > 0$.

(2) Suppose that rank $(V'_{(M')}(\underline{k})) = r \leq M'$. After some suitable equivalence transformations $V'_{(M')}(\underline{k})$ can be written as

$$\begin{bmatrix} V^1_{(M')}(\underline{k}) & V^2_{(M')}(\underline{k}) \\ V^3_{(M')}(\underline{k}) & V^4_{(M')}(\underline{k}) \end{bmatrix},$$

where $V^1_{(M')}(\underline{k})$ is $r \times r$ nonsingular matrix. But $V^1_{(M')}(\underline{k})$ can be enlarged to the size of $M' \times M'$ which is a Vandermode matrix with the following structure:

$$V^1_{(M')}(\underline{k}) = \begin{bmatrix} 1 & k_1 & k_1^2 & \cdots & k_1^{M'-1} \\ 1 & k_2 & k_2^2 & \cdots & k_2^{M'-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & k_{M'} & k_{M'}^2 & \cdots & k_{M'}^{M'-1} \end{bmatrix}.$$

The matrix $V^1_{(M')}(\underline{k})$ can be built by any selection of M' rows from the M rows of $V_{(M')}(\underline{k})$. The Vandermonde matrix has the following determinant:

$$\det\left\{V^1_{(M')}(\underline{k})\right\} = \prod_{i>j}^{M'} (k_i - k_j) \neq 0.$$

Since all the k_i are distinct, then $V^1_{(M')}(\underline{k})$ is nonsingular, and therefore, $r = M'$. ∎

Next, we show that the compressing matrices A are not uniquely determined and study the properties of the set $\tilde{A}$.

LEMMA 2. *The set $\tilde{A}$ is convex.*

PROOF. Let $A_1, A_2 \in \tilde{A}$. Let $\hat{F}_1$ and $\hat{F}_2$, be the reconstructed image with respect to A_1 and A_2. The original image for both is F, and the minimal distortion is E_r. We define A_3 as

$$A_3 = \alpha A_1 + \beta A_2, \qquad \alpha \geq 0, \quad \beta \geq 0, \qquad \alpha + \beta = 1.$$

The reconstructed image with respect to A_3 is $\hat{F}_3$, where

$$\hat{F}_3 = \alpha \hat{F}_1 + \beta \hat{F}_2,$$

because of the linearity.

$$\left\|F - \hat{F}_3\right\| = \left\|F - \alpha\hat{F}_1 - \beta\hat{F}_2\right\| = \left\|\alpha\left(F - \hat{F}_1\right) + \beta\left(F - \hat{F}_2\right)\right\|$$
$$\leq \alpha\left\|F - \hat{F}_1\right\| + \beta\left\|F - \hat{F}_2\right\| = \alpha E_r + \beta E_r = E_r.$$

But E_r is the minimal distortion. Therefore, A_3 is also best compressor and $A_3 \in \tilde{A}$. ∎

LEMMA 3. *Assume that $\hat{F}_1$ and $\hat{F}_2$ are two reconstructed images. If $\|F - \hat{F}_1\| \leq r$ and $\|F - \hat{F}_2\| \leq r$, then $\|2F - (\hat{F}_1 + \hat{F}_2)\| < 2r$, unless $\hat{F}_1 = \hat{F}_2$.*

PROOF. This is the property of strict convexity. The space of $M \times N$ matrices with the Euclidian Norm is strictly convex [13, p. 141]. ∎

On the other hand, the reconstructed image given the order of compression is unique.

LEMMA 4. *There is only one best reconstructed image $\hat{F}$, for all the elements of $\tilde{A}$.*

PROOF. Suppose there are two distinct best reconstructed images $\hat{F}_1$ and $\hat{F}_2$, for $A_1, A_2 \in \tilde{A}$. Then $\|F - \hat{F}_1\| = \|F - \hat{F}_2\| = Er(M', N')$, where $\mathrm{Er}(M', N')$ is minimal.

Now $F - \hat{F}_1$ and $F - \hat{F}_2$ are also distinct. We use Lemma 3 and $\|(F - \hat{F}_1) + (F - \hat{F}_2)\| < 2Er(M', N')$. Hence, $\|F - 1/2(\hat{F}_1 + \hat{F}_2)\| < Er(M', N')$. But this is a contradiction to the minimality of $\mathrm{Er}(M', N')$ (see [13]). ∎

LEMMA 5. *Let* $A_1, A_2 \in \tilde{A}$. *Then*

(1)
$$V'_{(M')}(\underline{k})(A_1 - A_2)V'_{(N')}(\underline{l})^{\top} = 0,$$

(2)
$$A_1 - A_2 = Y - V'_{(M')}(\underline{k})^{\perp} V'_{(M')}(\underline{k}) Y V'_{(N')}(\underline{l})^{\top} V'_{(N')}(\underline{l})^{\top^{\perp}},$$

where $V'^{\perp}$ *is the pseudo inverse of* V', *i.e.,* $V'V'^{\perp}V' = V'$, *and* Y *is some arbitrary* $M' \times N'$ *matrix.*

PROOF. (1) is derived directly from Lemma 4. By applying (1) and the property of the "pseudo inverse" we obtain

$$V'_{(M')}(\underline{k})\left[Y - V'_{(M')}(\underline{k})^{\perp} V'_{(M')}(\underline{k}) Y V'_{(N')}(\underline{l})^{\top} V'_{(N')}(\underline{l})^{\top^{\perp}}\right] V'_{(N')}(\underline{l})^{\top} = 0,$$

for every $M' \times N'$ Y matrix. Hence (2). ∎

LEMMA 6. *A necessary and sufficient condition for zero-error compression (lossless compression) is:*

$$F = V'_{(M')}(\underline{k}) V'_{(M')}(\underline{k})^{\perp} F V'_{(N')}(\underline{l})^{\top^{\perp}} V'_{(N')}(\underline{l})^{\top},$$

and the general solution is:

$$A = V'_{(M')}(\underline{k})^{\perp} F V'_{(N')}(\underline{l})^{\top^{\perp}} + Y - V'_{(M')}(\underline{k})^{\perp} V'_{(M')}(\underline{k}) Y V'_{(N')}(\underline{l})^{\top} V'_{(N')}(\underline{l})^{\top^{\perp}},$$

where Y *is* $M' \times N'$ *arbitrary matrix.*

PROOF. When there is no error:

$$\begin{aligned} F = V'_{(M')}(\underline{k}) A V'_{(N')}(\underline{l})^{\top} &= V'_{(M')}(\underline{k}) V'_{(M')}(\underline{k})^{\perp} V'_{(M')}(\underline{k}) A V'_{(N')}(\underline{l})^{\top} V'_{(N')}(\underline{l})^{\top^{\perp}} V'_{(N')}(\underline{l})^{\top} \\ &= V'_{(M')}(\underline{k}) V'_{(M')}(\underline{k})^{\perp} F V'_{(N')}(\underline{l})^{\top^{\perp}} V'_{(N')}(\underline{l})^{\top}. \end{aligned}$$

Conversely, if the general solution holds, then $A = V'_{(M')}(\underline{k})^{\perp} F V'_{(N')}(\underline{l})^{\top^{\perp}}$ is a particular solution. The generality of the solution follows since:

$$V'_{(M')}(\underline{k})\left[Y - V'_{(M')}(\underline{k})^{\perp} V'_{(M')}(\underline{k}) Y V'_{(N')}(\underline{l})^{\top} V'_{(N')}(\underline{l})^{\top^{\perp}}\right] V'_{(N')}(\underline{l})^{\top} = 0. \quad ∎$$

The main result is the extension and correction of [7,10], by showing the best least-square approximation, the set of all the "compressors" that construct the convex set $\tilde{A}$, and the expression for the error. We still do not know how to choose the minimal entropy element in the set $\tilde{A}$. The hypothesis that the minimal absolute value element A_0 is the minimal entropy element has not yet been justified rigoursly.

THEOREM 1.

(1) *Every "compressor" matrix* $A \in \tilde{A}(M', N')$ *can be described as:*

$$A = V'_{(M')}(\underline{k})^{\perp} F V'_{(N')}(\underline{l})^{\top^{\perp}} + Y - V'_{(M')}(\underline{k})^{\perp} V'_{(M')}(\underline{k}) Y V'_{(N')}(\underline{l})^{\top} V'_{(N')}(\underline{l})^{\top^{\perp}},$$

where Y *is an* $M' \times N'$ *arbitrary matrix.*

(2) *The minimal absolute value element* A_0 *is:*

$$A_0 = V'_{(M')}(\underline{k})^{\perp} F V'_{(N')}(\underline{l})^{\top^{\perp}}.$$

(3) *The minimal error* $Er(M', N')$ *for fixed* (M', N') *is:*

$$Er = \left\| \left(I - V'_{(M')}(\underline{k}) V'_{(M')}(\underline{k})^{\perp}\right) F \left(I - V'_{(N')}(\underline{l})^{\top^{\perp}} V'_{(N')}(\underline{l})^{\top}\right) \right\|.$$

PROOF. The decompression process at the receiver produces an approximated image represented in a matrix form as:

$$\hat{F} = V'_{(M')}(\underline{k})\, AV'_{(N')}(\underline{l})^{\top}. \tag{1}$$

From Lemma 4 we know that the optimal approximated image $\hat{F}$ is unique. From Lemma 5, we have that the contribution to the approximated image of the term

$$Y - V'_{(M')}(\underline{k})^{\perp} V'_{(M')}(\underline{k})\, YV'_{(N')}(\underline{l})^{\top} V'_{(N')}(\underline{l})^{\top^{\perp}} \tag{2}$$

is zero, after applying the decompression transform. Hence, the minimal distance (minimal distortion/error) is given by

$$d(A) = \left\|F - \hat{F}\right\| = \left\|F - V'_{(M')}(\underline{k})\, AV'_{(N')}(\underline{l})^{\top}\right\|. \tag{3}$$

We seek for the minimal matrix A of dimension $M' \times N'$, which minimizes the distance $d(A)$ between the original image F, and the reconstructed image $\hat{F}$. Hence, for any other matrix A' of the same size, either

$$\left\|F - V'_{(M')}(\underline{k})\, A'V'_{(N')}(\underline{l})^{\top}\right\| > \left\|F - V'_{(M')}(\underline{k})\, AV'_{(N')}(\underline{l})^{\top}\right\|, \tag{4}$$

or if A' belongs to the set $\tilde{A}$

$$\left\|F - V'_{(M')}(\underline{k})\, A'V'_{(N')}(\underline{l})^{\top}\right\| = \left\|F - V'_{(M')}(\underline{k})\, AV'_{(N')}(\underline{l})^{\top}\right\| \quad \text{and} \quad \|A'\| \geq \|A\|, \tag{5}$$

since

$$\begin{aligned}\min_{A}\left\|F - V'_{(M')}(\underline{k})AV'_{(N')}(\underline{l})^{\top}\right\| &= \min_{A}\left\|V'_{(M')}(\underline{k})AV'_{(N')}(\underline{l})^{\top}V'_{(N')}(\underline{l})^{\top^{\perp}} - FV'_{(N')}(\underline{l})^{\top^{\perp}}\right]V'_{(N')}(\underline{l})^{\top}\right\| \\ &= \min_{A}\left\{\left\|V'_{(M')}(\underline{k})\, A - FV'_{(N')}(\underline{l})^{\top^{\perp}}\right\|\left\|V'_{(N')}(\underline{l})^{\top}\right\|\right\}. \end{aligned} \tag{6}$$

Since $\|V'_{(N')}(\underline{l})^{\top}\|$ is positive and independent on A, then the minimization is done only on the first term. Denote:

$$V \triangleq V'_{(M')}(\underline{k}), \qquad B \triangleq FV'_{(N')}(\underline{l})^{\top^{\perp}}. \tag{7}$$

Then by using (7) the minimization problem (6) becomes

$$\min_{A}\|VA - B\|. \tag{8}$$

Using the results by Penrose [14,15] we can formulate:

$$V^{\top}VV^{\perp} = V^{\top}, \tag{9}$$

$$[VP + (I - VV^{\perp})Q]^{\top}[VP + (I - VV^{\perp})Q] = (VP)^{\top}VP + [(I - VV^{\perp})Q]^{\top}[(I - VV^{\perp})Q], \tag{10}$$

for suitably dimensioned matrices P and Q. That is

$$\left\|VP + (I - VV^{\perp})\, Q\right\|^2 = \|VP\|^2 + \left\|(I - VV^{\perp})\, Q\right\|^2. \tag{11}$$

In our particular case:

$$\begin{aligned}\|VA - B\|^2 &= \left\|V(A - V^{\perp}B) + (I - VV^{\perp})(-B)\right\|^2 \\ &= \left\|V(A - V^{\perp}B)\right\|^2 + \left\|(I - VV^{\perp})(-B)\right\|^2 \geq \left\|VV^{\perp}B - B\right\|^2, \end{aligned} \tag{12}$$

where equality occurs only when the first term vanishes, i.e., when $VA = VV^{\perp}B$. Replacing V by $V^{\perp}$ in (11) produces:

$$\left\|V^{\perp}B + \left(I - V^{\perp}V\right)A\right\|^2 = \left\|V^{\perp}B\right\|^2 + \left\|\left(I - V^{\perp}V\right)A\right\|^2. \tag{13}$$

Thus, if $VA = VV^{\perp}B$, then (13) gives:

$$\|A\|^2 = \left\|V^{\perp}B\right\|^2 + \left\|A - V^{\perp}B\right\|^2. \tag{14}$$

Thus, from (12) and (14), we see that if $A = V^{\perp}B$, the conditions of (4) and (5) are satisfied. Hence, substituting (7) we have

$$A = V'_{(M')}\left(\underline{k}\right)^{\perp} F V'_{(N')}\left(\underline{l}\right)^{\mathrm{T}^{\perp}}. \tag{15}$$

The minimal error Er is obtained when (15) holds:

$$\begin{aligned} Er &= \left\|F - \hat{F}\right\| = \left\|F - V'_{(M')}\left(\underline{k}\right)V'_{(M')}\left(\underline{k}\right)^{\perp} F V'_{(N')}\left(\underline{l}\right)^{\mathrm{T}^{\perp}} V'_{(N')}\left(\underline{l}\right)^{\mathrm{T}}\right\| \\ &= \left\|\left(I - V'_{(M')}\left(\underline{k}\right)V'_{(M')}\left(\underline{k}\right)^{\perp}\right) F \left(I - V'_{(N')}\left(\underline{l}\right)^{\mathrm{T}^{\perp}} V'_{(N')}\left(\underline{l}\right)^{\mathrm{T}}\right)\right\|. \end{aligned} \tag{16}$$

∎

COROLLARY 1. *The expression for the minimal error and especially the condition for zero-error, dictate the policy for best compression. The point (M', N') for best distortion-free compression is the point where the condition of Lemma 6 holds, and this condition does not hold for the points $(M'-1, N')$ and $(M', N'-1)$.*

3. IMAGE COMPRESSION WITH A COST CRITERION

A practical question in image compression in areas like storage, transmission, etc., is: what is the best target-transform in the sense of minimum cost? Obviously, the answer depends on the definition of the cost function. We define a cost function as a tradeoff between the number of coefficients that represent the image and the error. The user can adjust it based on the previous still images of the same type. The function J is

$$J = \alpha\left[\text{Min-Error}\right] + \beta\left[\text{Transmitted-Information}\right]. \tag{17}$$

We will treat the case of an image which is defined on a Cartesian grid as it is described below. As a consequence of Theorem 1, J depends only on $M' \times N'$

$$J\left(M', N'\right) = \alpha\left[\text{Min-Error}\right] + \beta\left[I_0\right] = \alpha\left[\text{Min-Error}\right]_{(M',N')} + \beta M'N'. \tag{18}$$

The ratio α/β will, certainly, determine some optimal point (M', N'), where J is minimized. The following properties of $J(M', N')$ are valid:

$$J\left(M', N'\right) \geq J\left(M'+1, N'\right), \tag{19a}$$

$$J\left(M', N'\right) \geq J\left(M', N'+1\right), \tag{19b}$$

$$J\left(M, N\right) = \beta MN. \tag{19c}$$

Denote by $d\left(M', N'\right)$ the [Min-Error] which is the minimum distance obtained from compression, to size $M' \times N'$-transform. The cost-function in this simple control problem depends only on "end-point" -(M', N'). By using partial derivative with respect to M' and N', we can obtain the optimal point (M', N'). By variation calculus $\delta J = 0$ at the optimal point (M', N').

$$\left[\alpha\frac{\partial d\left(M', N'\right)}{\partial M'} + \beta N'\right]\delta M' = 0, \tag{20a}$$

$$\left[\alpha\frac{\partial d\left(M', N'\right)}{\partial N'} + \beta M'\right]\delta N' = 0. \tag{20b}$$

Since M' and N' are integers, $\delta N' = \delta M' = 1$. The point (M', N') is determined near the zero-crossing point of

$$\alpha \frac{\partial d(M', N')}{\partial M'} + \beta N', \tag{21a}$$

$$\alpha \frac{\partial d(M', N')}{\partial N'} + \beta M'. \tag{21b}$$

Hence,

$$N'_{\text{opt}} \approx -\frac{\alpha}{\beta} \frac{\partial d(M', N')}{\partial M'} \Big|_{M'=M'_{\text{opt}}, N'=N'_{\text{opt}}}, \tag{22a}$$

$$M'_{\text{opt}} \approx -\frac{\alpha}{\beta} \frac{\partial d(M', N')}{\partial N'} \Big|_{M'=M'_{\text{opt}}, N'=N'_{\text{opt}}}. \tag{22b}$$

The ratio α/β determines the optimal compression-point. Equations (22a) and (22b) should be solved numerically by using the definition (16) of the minimal distance $d(M', N')$. The calculations are done on a specific grid and a typical image F.

4. IMAGE COMPRESSION IN A FIXED ERROR

If some LSE distortion D is tolerable in image reconstruction, then the optimal compression point M', N' is derived from the distortion function. The distortion $d(M', N')$ is a function given in (15) of (M', N') on a specific grid and a typical image F. Therefore, determination of a tolerable distortion D defines a region of acceptable points bounded by the curve $d(M', N') = D$.

5. EXPERIMENTAL RESULTS

As indicated in the introduction of this paper, we focus on theory rather than practice—the reader interested in recent results in image coding by polynomial approximation can profitably consult the survey of Kunt, Benard and Leonardi [7]. We have not conducted a full scale research of image encoding by using adaptive split-and-merge coding methods. However, we have addressed practical issues encountered in lossy polynomial encoding of images such as: quantization, bit allocation, segmentation, pyramidal decomposition, variance of coefficients, and error histograms. In the following section we give some experimental results supporting our developments and emphasizing on the open problems. As mentioned above, a split-and-merge algorithm is essential as a preprocessing stage. But this is beyond the scope of this paper, which is solely dedicated to polynomial approximation. Certainly, there is a place for further research in the field, and we highlight the main issues.

In our work performed by Shkliarman and Avrin, there was no adaptive split stage. The image was split to squared subimages of 8×8 pixels in each such subimage. We used the minimal absolute value transform (15) for compression to a 2×2 coefficients matrix.

The chosen image was "Lena," presented in Figure 3, even though it was a difficult challenge to overcome the obstacles of contours in a human face. This image contains a mixture of lines, edges, and other types of contours. It is desirable to use first a contour detector that responds appropriately to the various contour types. However, in making any comparisons to [7] or others, we must keep in mind that our scheme has not involved a contour detecting stage.

In Figures 4–8, we present several choices of quantization levels for the coefficients of the transform-matrix. After output from the transform-matrix, each of the coefficients is uniformly quantized. The purpose of quantization is to achieve further compression by representing the coefficients with no greater precision than is necessary to achieve the desired image quality. Stated another way, the goal of this processing step is to discard information which is not visually

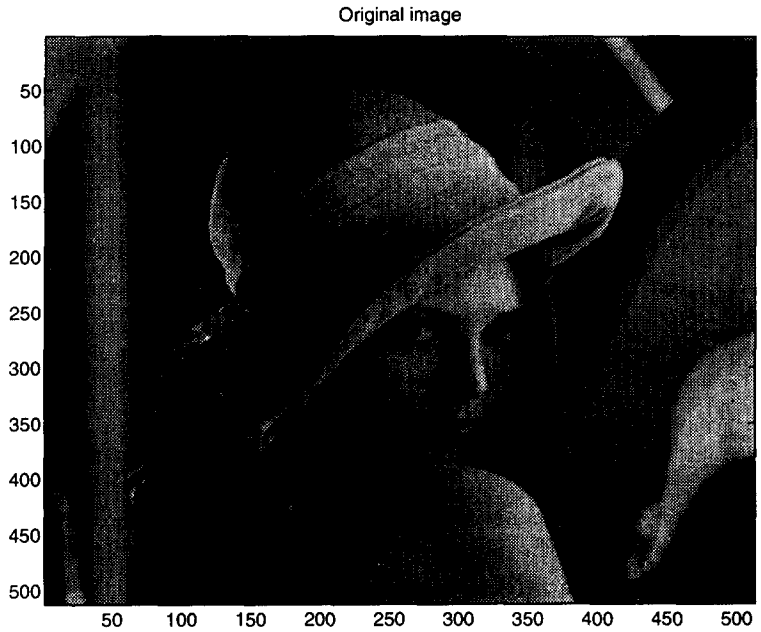

Figure 3. "Lena," the original image.

Figure 4. Quantization of coefficients results.

significant. Quantization is a many-to-one mapping, and therefore, is fundamentally lossy. It is one of the sources of lossiness in the encoder. The different allocation of the levels causes different values for the bit-rate and the accumulated square error. The dependence of bit-rate (a function of coefficients quantization) versus error is presented in Figure 9 for the transformation from an 8×8 subimage, to a 2×2 transform-matrix. We have observed that increasing the bit-rate reduces the blocking effect and the distortion. Empirically, it is clear that the best approximation is obtained in the low-pass area of the image, and the errors are concentrated around the high-pass regions—the contours.

Later, we studied the variance of the coefficients in a 4×4 coefficients matrix. Figure 10 shows that the low order coefficients are much more correlated than the higher order coefficients, due to the low-pass property of most of the image. This result supports the basic motivation to segment

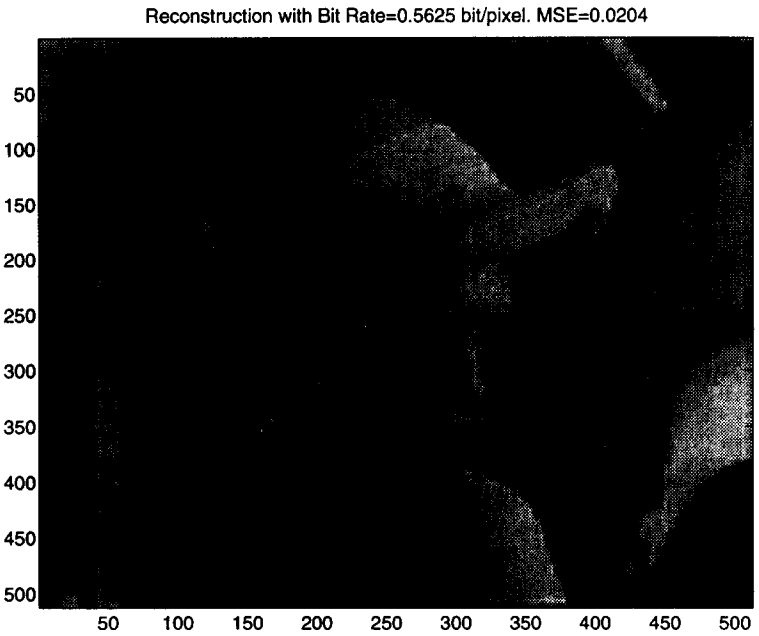

Figure 5. Coefficients quantization results.

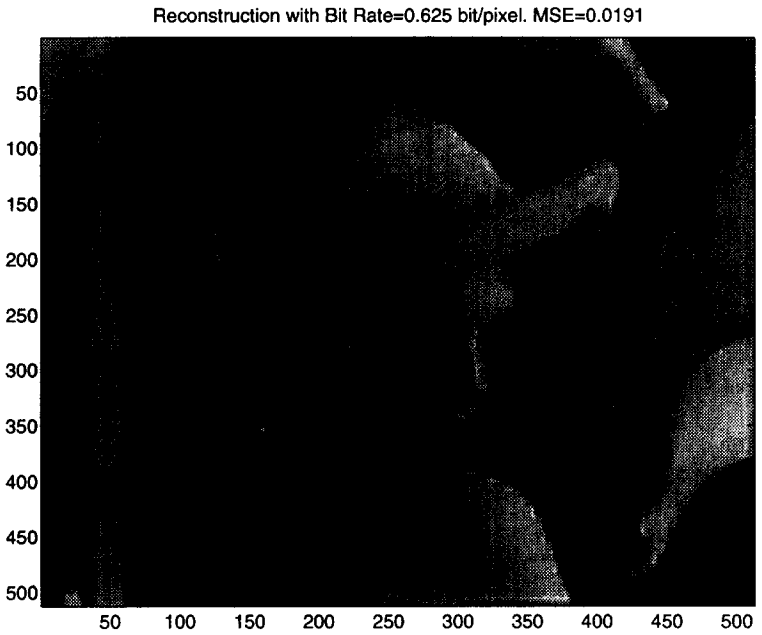

Figure 6. Coefficients quantization results.

the image to low-pass regions and then perform polynomial approximation. Such regions are very well approximated by a low order 2-D polynomial.

Figure 11 presents a pyramidal structure of the image. Two stages are shown, where the second picture is obtained by successive decomposition from the first one. We used 8×8 subimages transformed by 8×8 matrix-transform in both stages.

Figure 12 presents a reconstructed image after compression where the coefficients are not quantized and are represented by floating point. Subimages of 8×8 size have been transformed to 2×2 matrices.

Figure 13 presents a pyramidal structure of the image. Two stages are shown, where the second picture is obtained by successive decomposition from the first one. We used 8×8 subimages transformed by a 4×4 matrix-transform in both stages.

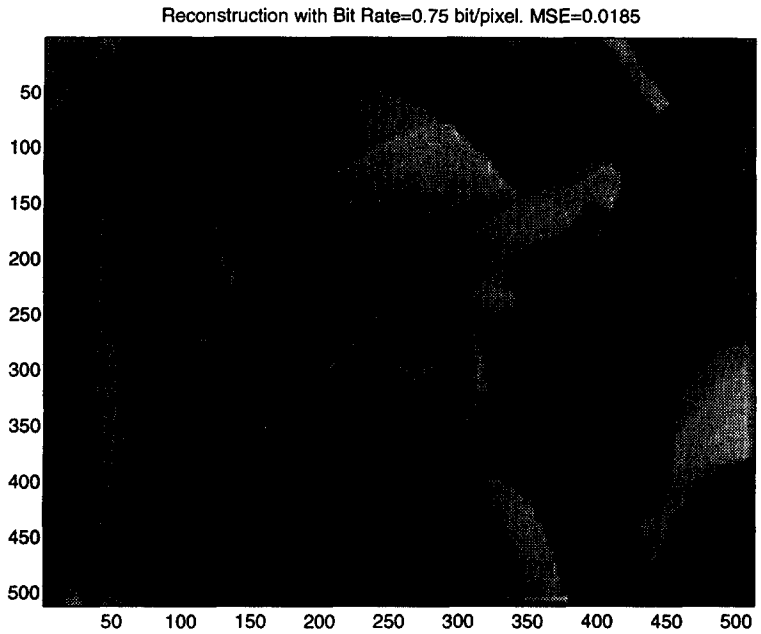

Figure 7. Quantization results.

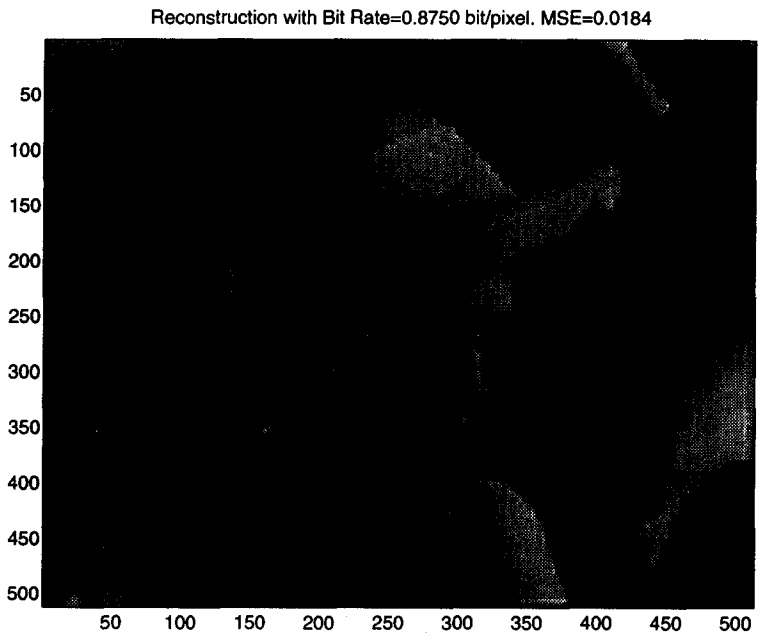

Figure 8. Quantization results.

Figures 14 and 15 present reconstructed images after pyramidal compression where the coefficients are not quantized and represented by floating point. Subimages of 8×8 size have been transformed by 2×2 matrices.

The histograms of the Error Image between the reconstructed image and the original image, and the second degree error result in a Gaussian-like function, concentrated around zero. The graphs are plotted in Figures 16 and 17. The obtained entropies are 0.2774 for the first order Error Image, and 1.0339 for the second order Error.

We consider these results as preliminary. We pose the question of how to search for the best compressor among all the members in the optimal set. The crux of the matter is the minimal entropy selection of the transform-matrix.

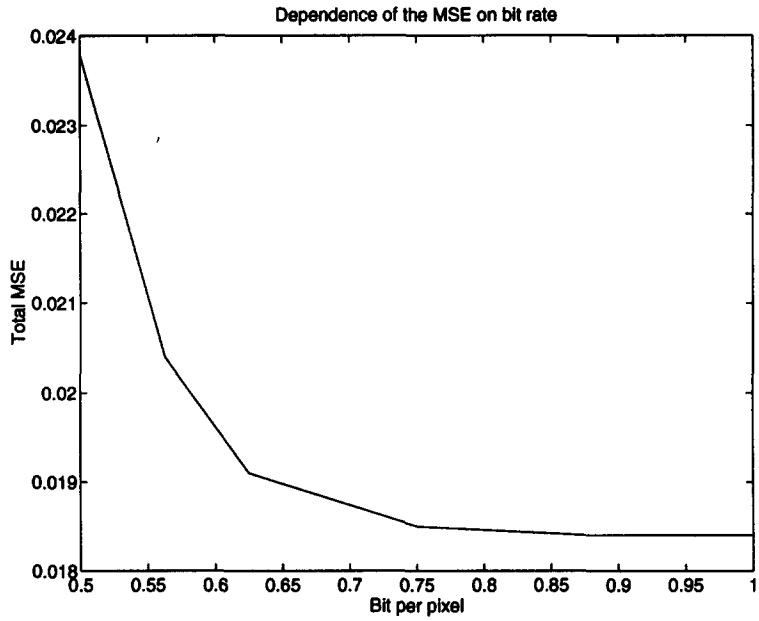

Figure 9. Dependence of MSE on quantization of coefficients.

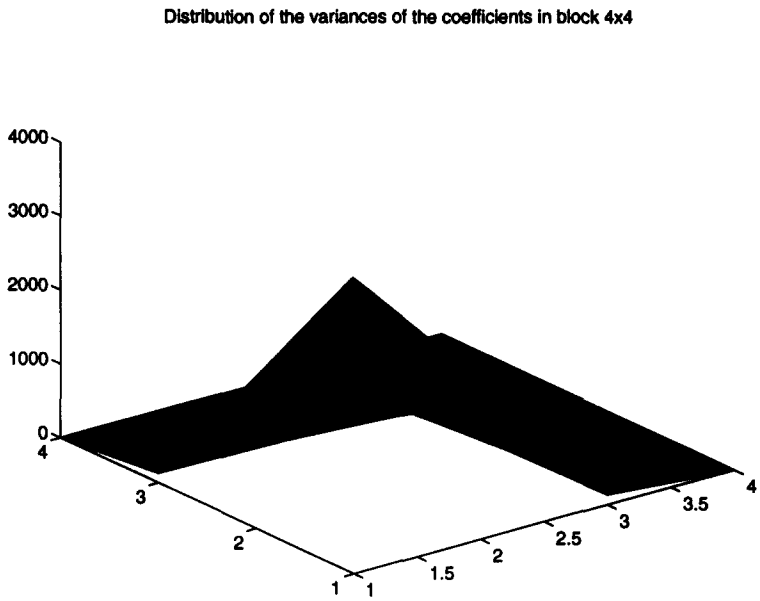

Figure 10. Distribution of coefficients variance.

6. CONCLUSIONS

In this paper, a brief overview is first given of image coding techniques that attempt to reach high compression using segmentation by adaptive split-and-merge procedures. We have concentrated only on the approximation stage that follows the segmentation. A good segmentation allows us to use polynomials of reduced order to approximate image data over large regions.

Next, we propose to compress the image by building a pyramid-style description where the low-pass operation at each stage is a polynomial approximation. The image data is a slowly varying luminance function over the region, assuming that a good segmentation has been performed. Such variations are well represented by 2-D polynomials. Moreover, the oscillatory behavior of commonly used orthogonal functions does not exist for polynomial approximation.

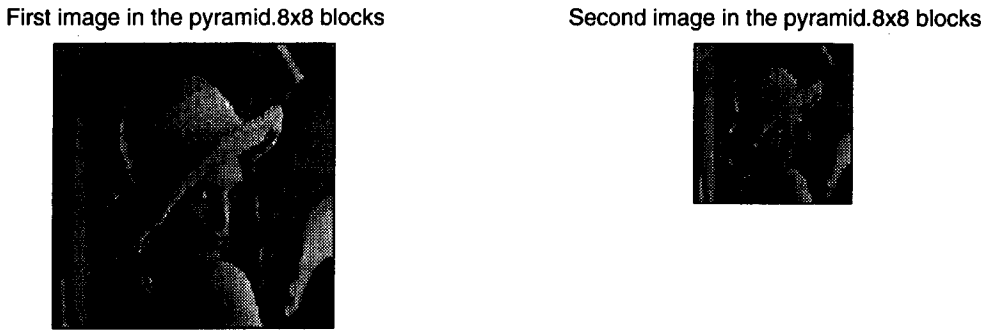

Figure 11. Pyramidal structure.

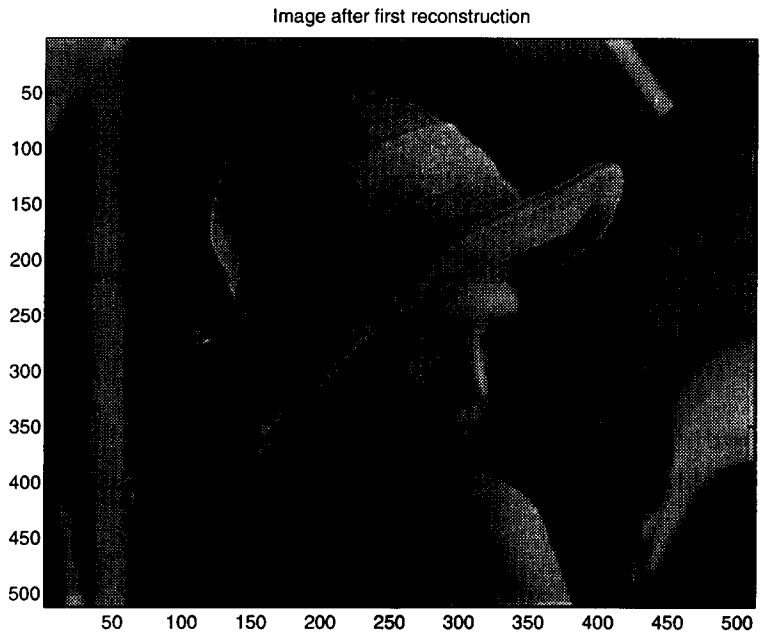

Figure 12. Reconstruction based on pyramidal structure.

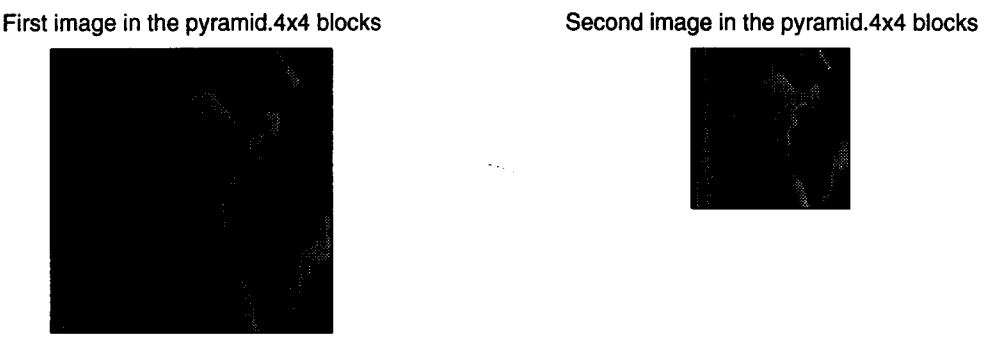

Figure 13. Pyramidal decomposition.

Following the outlined motivations for choosing that kind of approximation, and definitions of cost functions, and optimality, we study in details the properties of the transform obtained by 2-D polynomial approximation.

The main results include the following:

(1) For a given compression ratio in the two-dimensions, and a given LSE distortion level, the reconstructed image is unique.

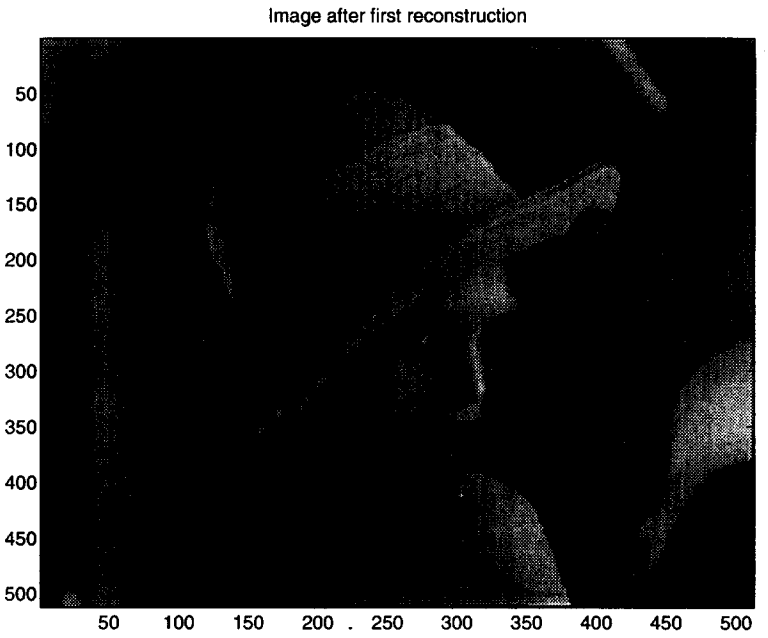

Figure 14. Reconstruction based on pyramidal structure.

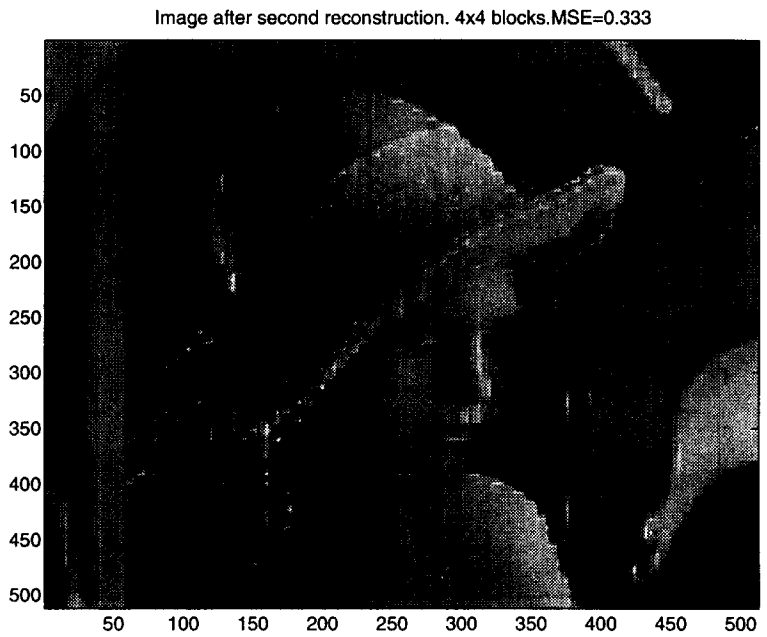

Figure 15. Reconstruction of pyramidal compressed image.

(2) There are infinitely many possible solutions for the best compressors and they compose a convex set. However, it is still an open question if the minimal absolute value element in the set is the best compressor in the sense of minimum entropy.
(3) We present a method for calculating all the members in the set of optimal compressors.
(4) Calculation of the compressors is involved with "reduced Vandermonde matrices." It is shown that these are full rank positive definite matrices.
(5) The conditions for lossless compression are proved.
(6) The expression of the error caused by compression is proved and used later for determination of compression ratio with a fixed error.
(7) A tradeoff between error and compression ratio, leads to a new definition of a cost function. Image compression subject to such criterion is discussed.

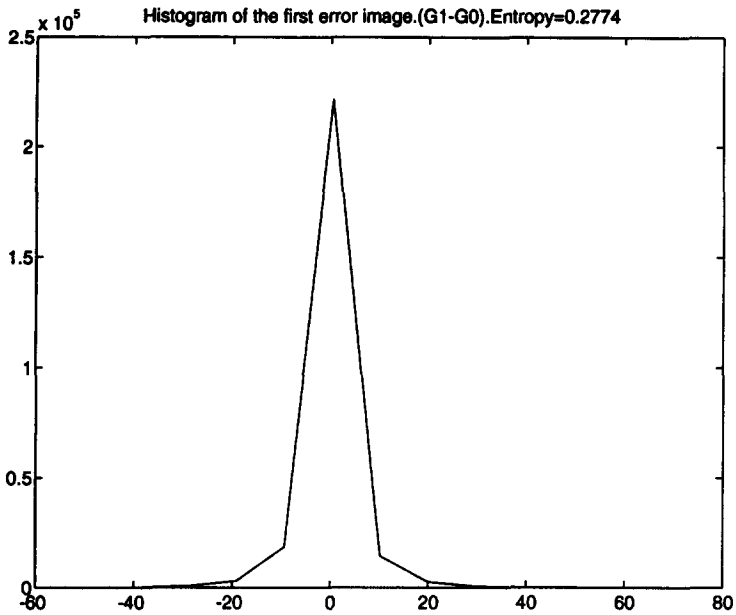

Figure 16. Distribution of the error.

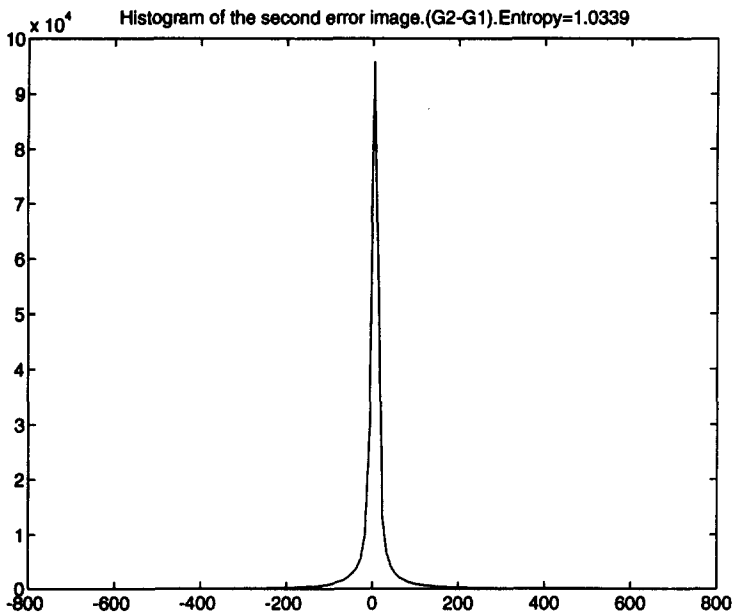

Figure 17. Distribution of second order error.

The proposed continuous-tone image compression method is not a panacea that will solve the myriad issues, which must be addressed before digital images can be fully integrated within all the applications that will ultimately benefit from them. However, the results of [7,13], and our work are expected to yield a method that will withstand the tests of quality and time, where methods such as JPEG fail. The oscillatory behavior of orthogonal functions such as DCT, do not exist for polynomials.

REFERENCES

1. C.E. Shannon, Coding theorems for a discrete source with a fidelity criterion, *IRE Nat. Conv. Rec. Part 4*, 142–163, (1959).
2. T. Berger, *Rate Distortion Theory*, Prentice Hall, Englewood Cliffs, NJ, (1971).

3. R.M. Gray, *Source Coding Theory*, Kluwer, Boston, MA, (1990).
4. J.C. Kieffer, A survey of the theory of source coding with a fidelity criterion, *IEEE Trans. on Information Theory* **IT-39** (5), (September 1993).
5. H. Abut, *Vector Quantization*, IEEE Press, New York, (1990).
6. N.S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice Hall, Englewood Cliffs, NJ, (1984).
7. M. Kunt, M. Benard and R. Leonardi, Recent results in high compression image coding, *IEEE Trans. on Circuits and Systems* **34** (11), 1306–1336, (November 1987).
8. A.N. Netravali and J.O. Limb, Picture coding: A review, *Proc. IEEE* **68**, 336–406, (1980).
9. W.K. Pratt, Editor, *Image Transmission Techniques*, Academic Press, New York, (1979).
10. M. Eden, M. Unser and R. Leonardi, Polynomial representation of pictures, *Signal Processing* **10**, 385–393, (1986).
11. P.J. Burt and E.H. Adelson, The Laplacian pyramid as a compact image code, *IEEE Trans. on Communciations* **COM-31** (4), (April 1983).
12. S.G. Mallat, A theory for multiresolution signal decomposition: The wavelet representation, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **PAMI-11** (7), (July 1989).
13. P.J. Davis, *Interpolation and Approximation*, Dover, (1975).
14. R. Penrose, On best approximate solutions of linear matrix equations, *Proc. Cambridge Philos. Soc.* **52**, 17–19, (1956).
15. R. Penrose, A generalized inverse matrices, *Proc. Cambridge Philos. Soc.* **51**, 406–413, (1955).